Canada, Statistics Canada

GRDSR: Facts by small areas

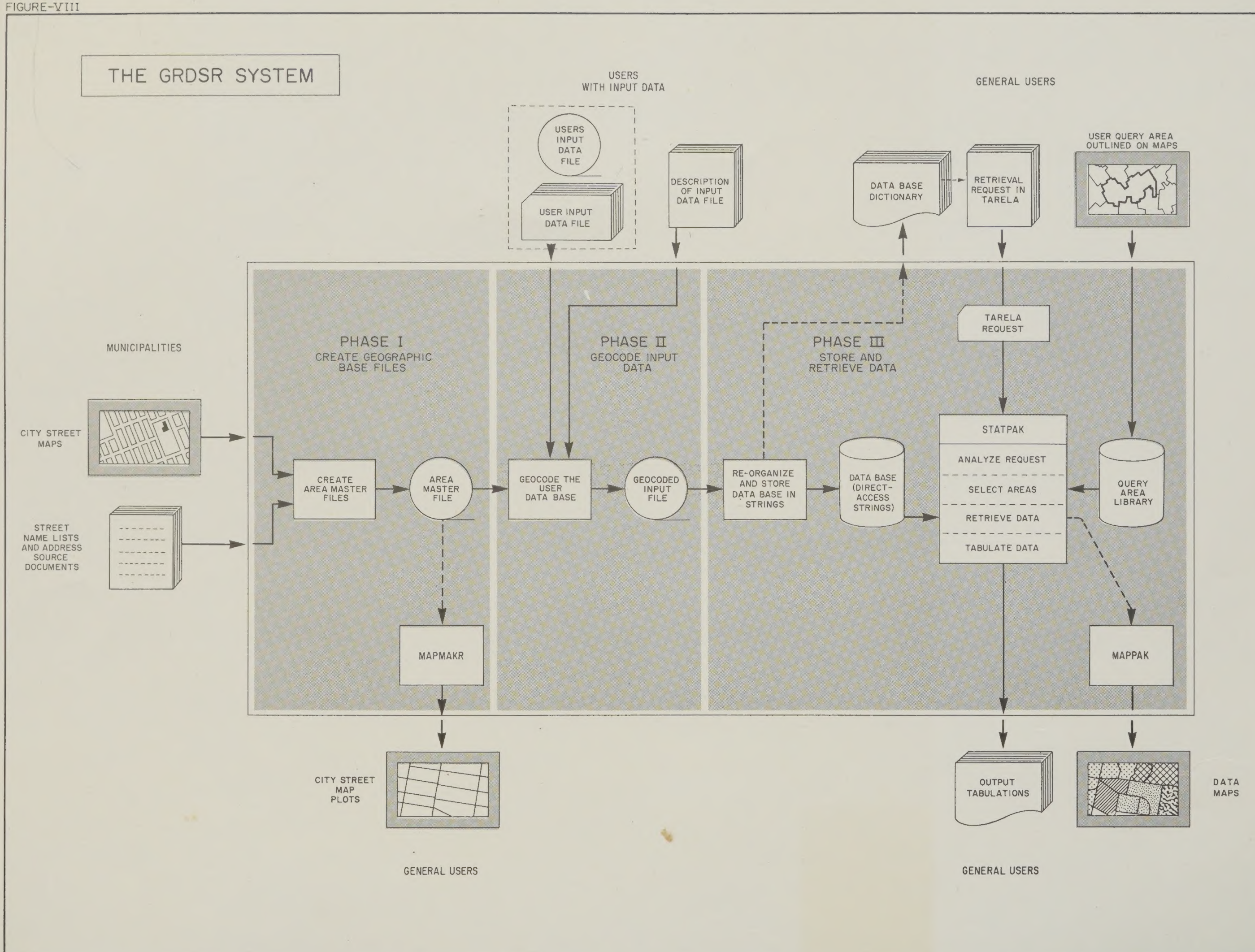
The Geographically Referenced Data Storage and Retrieval System

AN INTRODUCTION

JUNE 1972

Statistics Canada Statistique Canada

FIGURE-VIII


THE GRDSR SYSTEM
USERS WITH INPUT DATA
GENERAL USERS
USERS INPUT DATA FILE
USER INPUT DATA FILE
DESCRIPTION OF INPUT DATA FILE
DATA BASE DICTIONARY
RETRIEVAL REQUEST IN TARELA
USER QUERY AREA OUTLINED ON MAPS
MUNICIPALITIES
CITY STREET MAPS
STREET NAME LISTS AND ADDRESS SOURCE DOCUMENTS
PHASE I
CREATE GEOGRAPHIC BASE FILES
PHASE II
GEOCODE INPUT DATA
PHASE III
STORE AND RETRIEVE DATA
TARELA REQUEST
CREATE AREA MASTER FILES
AREA MASTER FILE
GEOCODE THE USER DATA BASE
GEOCODED INPUT FILE
RE-ORGANIZE AND STORE DATA BASE IN STRINGS
DATA BASE (DIRECT-ACCESS STRINGS)
STATPAK
ANALYZE REQUEST
SELECT AREAS
RETRIEVE DATA
TABULATE DATA
QUERY AREA LIBRARY
MAPMAKR
MAPPAK
CITY STREET MAP PLOTS
OUTPUT TABULATIONS
DATA MAPS
GENERAL USERS
GENERAL USERS

GRDSR
(The geographically referenced data storage and retrieval system)

A new method of assembling statistical information by user-specified areas

AN INTRODUCTION

Published by Authority
of the Minister of Industry, Trade and Commerce

June 1972

CONTENTS

INTRODUCTION

What is GRDSR?

A unique and flexible system now makes it possible, for the first time, to provide information by user-specified areas in Canada's larger urban centres. Fully computerized, the GRDSR (Geographically Referenced Data Storage and Retrieval) system is the outcome of five years' research by Statistics Canada into solving the many problems associated with the storage and retrieval of statistics about small areas.

Through GRDSR, statistical information can now be quickly and inexpensively obtained about retrieval areas that range in size from a few city blocks to an entire urban centre.

Retrieval is made possible through a technique called geocoding, whereby urban areas are divided into many small building blocks or micro-areas. The blocks must be small enough that they can be assembled to approximate most retrieval areas required by users. Each building block is assigned a unique identifying coordinate number which, in turn, allows files of households, persons, or events to be coded to appropriate building blocks in the city area. The appropriate building block is usually the place of residence or location where the event occurred. At this point, the files are said to be geocoded. When an interested user needs information from a geocoded file, he outlines his area of interest (or "query area") on a map and makes a request. GRDSR then identifies all the building blocks contained within his query area and, using the corresponding coordinates, automatically retrieves all data belonging to the blocks. The statistics are then tabulated in a convenient report.

Background to GRDSR development

The need for concise, timely statistics is constantly pressing in many sectors of the economy. Diverse planning and decision-making efforts are often frustrated by a lack of relevant and timely data; the socio-economic benefits of more fully informed decisions may, as a result, be diluted or lost. Today, it is obvious that the pressure for diverse and specialized statistical information cannot help but increase. The comprehensive and thorough use of data already collected is now, therefore, more relevant than ever.

At Statistics Canada, this pressure is evident not only in the mounting volume of special data requests but by their changing nature. The trend consistently points to the need for flexible information systems fully capable of retrieving data on a specialized, often one-time basis. The basic requirement is, essentially, for an integrated information service — not just a data collection facility.

The development of GRDSR thus focuses upon an important trend, and the nature of this trend is clear: user requirements will increase, in terms of the amount of data required, types of aggregation and manipulation available, ease of retrieval, the format of the final statistical product and, of course, the response time. The evolution of the GRDSR system is now at the point where each of these requirements is substantially met.

Nature of the Problem

The gathering of small-area data has presented a difficult problem for some time. Urban planners, municipal agencies, school administrations and governments each impose different zoning patterns or jurisdictions over settled areas of land. Many agencies maintain records and use their own jurisdictions to collect and identify statistical information. At some point these records may attract general interest. But problems arise when outside groups try to use this information, because their requirement is for facts related to different area breakdowns.

Today, special-interest areas such as marketing zones, census tracts, school districts and land-use areas are in everyday use in major cities. However, these areas usually overlap and have little in common but the land area they reference. Thus, it is difficult to relate information from one source to outside areas of interest (see Figure I, page 2.)

In the past, when the sole means of disseminating statistics from the census was through published volumes, the statistics had to be summarized in terms of enumeration areas, census tracts or other standard areas. The standard census areas did not, however, coincide with many query areas for which data were required. Consequently, the requirements of many census data users could either not be met, or met only with great difficulty, at considerable cost and with considerable delay.

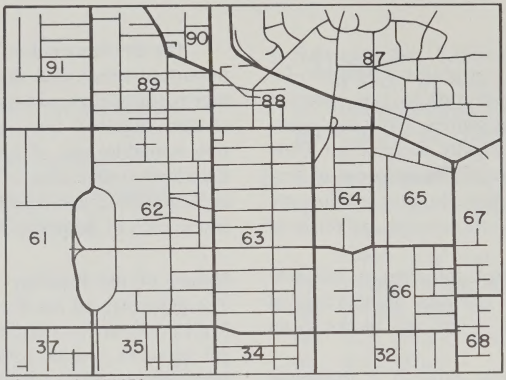
Where can GRDSR be Applied?

Given that the fundamental purpose of GRDSR is to allow users more flexibility in obtaining information about special-purpose areas, it is significant that the first major application of GRDSR has been the 1971 Census of Canada.

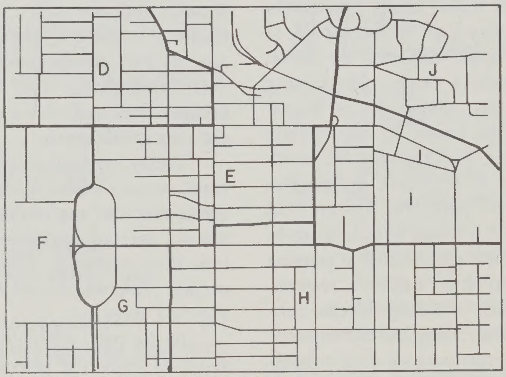
Originally conceived in anticipation of special census requests, the system has since been developed for general-purpose applications. Municipal assessment files, fire and accident reports, marketing surveys and hospital records are among several applications discussed in the next section.

FIGURE - I

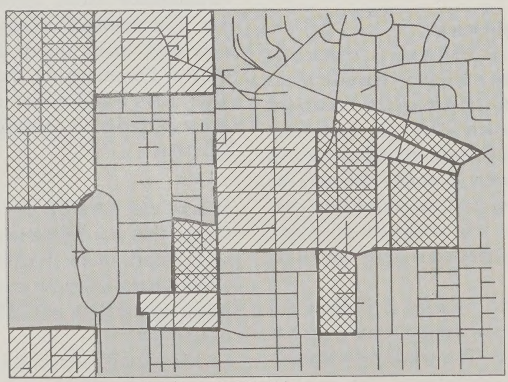
HOW USERS IMPOSE DIFFERENT ZONING PATTERNS OVER A CITY AREA



(1) CENSUS TRACTS



(2) SCHOOL DISTRICTS (HYPOTHETICAL)



(3) PLANNING ZONES (HYPOTHETICAL)

RESIDENTIAL OR PARKLAND

COMMERCIAL

INDUSTRIAL

APPLICATIONS AND POTENTIAL OF GRDSR

While Statistics Canada expects to serve many requests for statistical information from geocoded 1971 Census data, GRDSR is designed to handle the majority of address-bearing files and survey data which originate in larger urban centres in Canada (see Figure II, page 5.) Extensive geocoding applications are now possible in both the public and private sectors. Potential users include municipalities, planning and research groups, industrial and commercial firms, public utilities, social agencies, universities and governments — in short, any group using geographically-based information for research, planning or decision-making.

Noteworthy features of the Census application are outlined in part (i). Next, in part (ii), a number of other specific GRDSR applications are discussed. Finally, some aspects of a possible geocoding application, health services planning, are described in part (iii).

The Geocoded Census

GRDSR will provide a new dimension in census retrieval services: the facility to provide statistical data for user-specified areas anywhere in Canada.

Confidentiality

While Statistics Canada attaches great importance to meeting the need for custom-made, user-oriented data on a uniform, national basis, it can only do so within the confidentiality constraints imposed by the Statistics Act (1971). As a result, no information can be disseminated in such a way as to identify an individual respondent. Automatic routines within the system ensure that no such disclosure of information is possible.

Query Areas

In 14 larger urban centres (see Figure II, page 5), users may request data for areas as small as a few city blocks. Users should not, however, expect to receive representative data for smaller areas, such as one side of a block. There are two important reasons:

First, the results would be subject to high response and sampling errors, due to the small number of cases on which the statistics would be based. The usual process of compensating errors for larger samples could only take place to a limited extent.

Second, a carefully-controlled amount of statistical error is purposely introduced to all retrieved data so that no census respondent can be identified from the final tabulations. This random error, while of little or no significance to normal tabulations, would further obscure any information obtained about very small areas.

Outside the major urban centres, statistical information will be available at a coarser level of geographical detail. Here, query areas will be assembled using traditional census enumeration areas (EA's), which contain approximately 150-200 households each. As a result, extensive census data will be available for more than 27,000 EA's — either individually, or in any aggregation of interest to the user.

In either case, the desired areas are simply outlined on a suitable map, named clearly and submitted to Statistics Canada along with the tabulation request.

Request Language

Users may request census tabulations using an English-like language called TARELA (Tabulation Request Language), which can be learned in a few hours without previous programming knowledge. TARELA allows subject-matter specialists to write requests in terms familiar to their work. With this language users can create cross-tabulations of any combination of 1971 Census variables (which number more than 120) and generate tables having up to 10 dimensions. Users who are not familiar with TARELA can, of course, submit their request in precise narrative form or in the form of "dummy" tables. The required TARELA coding will then be generated by Statistics Canada.

Data Mapping

In addition to supplying census data in tabular form, GRDSR also includes a facility for data mapping. MAPPAK, which incorporates the Harvard mapping package SYMAP, is a remarkable feature in that it can accurately depict the distribution of data values over any area in graphical form. This type of map is particularly useful in locating areas where extreme values of some factor occur, and can be used to reveal problem areas at a glance.

Both TARELA and MAPPAK are general-purpose features of GRDSR, by no means limited to census applications. TARELA and MAPPAK are further described in Features and Components, page 10.

General Applications

Geocoding applications can be served by many data bases in addition to the Census. The system is designed to geocode many types of address-identified files, provided they originate in one of the larger Canadian urban centres (See Figure II, page 5.) The GRDSR programs will be available to municipal users wishing to geocode local data bases. Examples of suitable data bases include assessment rolls, traffic surveys, hospital

and welfare records, marketing surveys, school census data and certain accident, fire and police records.

Municipal Administration and Government

Public Services: Research studies are being conducted, using geocoding, to determine the frequency of accident, fire and police reports originating from various sections of large cities. Such statistics would clearly be a significant aid in planning or re-allocating municipal resources and services; the use of GRDSR is possible whenever records of such incidents are address-identified.

Education: A new method for planning the location of new schools and school districts is now possible through geocoding. Facts related to this application may include the concentrations and age distribution and projected growth rates of school-aged children within the community.

The routing of school buses is another application where geocoding offers considerable promise. GRDSR is ideally suited to provide statistics such as the geographic distribution of school-aged children.

Other applications include analysis of districts by such socio-economic factors as country of origin, language, religion, occupation and income as an input to planning of day-school curriculums and adult-education programs.

Urban Planning

Interests in the urban planning area include study and analysis of planning zones, optimizing the location of city services and facilities, planning of mass transit and analysis of potential urban renewal areas, land values and housing data.

In planning the route of a new city transit system, for instance, the starting points and destinations of potential users form a definite network or pattern. Subject to further analysis, such as transportation modelling, this network can have decisive impact on the final route chosen.

Further possibilities include planning of municipal services according to socio-economic factors such as population density, language, and income within selected urban zones. New approaches to planning the nature and location of welfare services may also become possible.

Medical Services

Typical problems include planning the location of hospitals, out-patient clinics and medical centres, and the establishment of a geographically-referenced inventory of nurses.

Industry, Commerce and Utilities

Geocoding has played a part in the allocation of facilities and services such as telephone exchanges and banks. Other applications include population and demographic studies of city areas, the planning of marketing zones and radio and television coverages, the optimization of retail store location in terms of customer proximity and resource allocation problems faced by oil, hydro and gas utilities.

A number of simulation and modelling techniques exist for solving network problems in the commercial transportation/distribution area. Typically, data related to some grid pattern constitute an essential requirement for this approach. GRDSR is an ideal research tool to help meet this need.

Universities

Interests include economic, political and social studies of neighbourhoods, electoral districts and socio-economic research into city areas defined by such factors as country of origin, language or income.

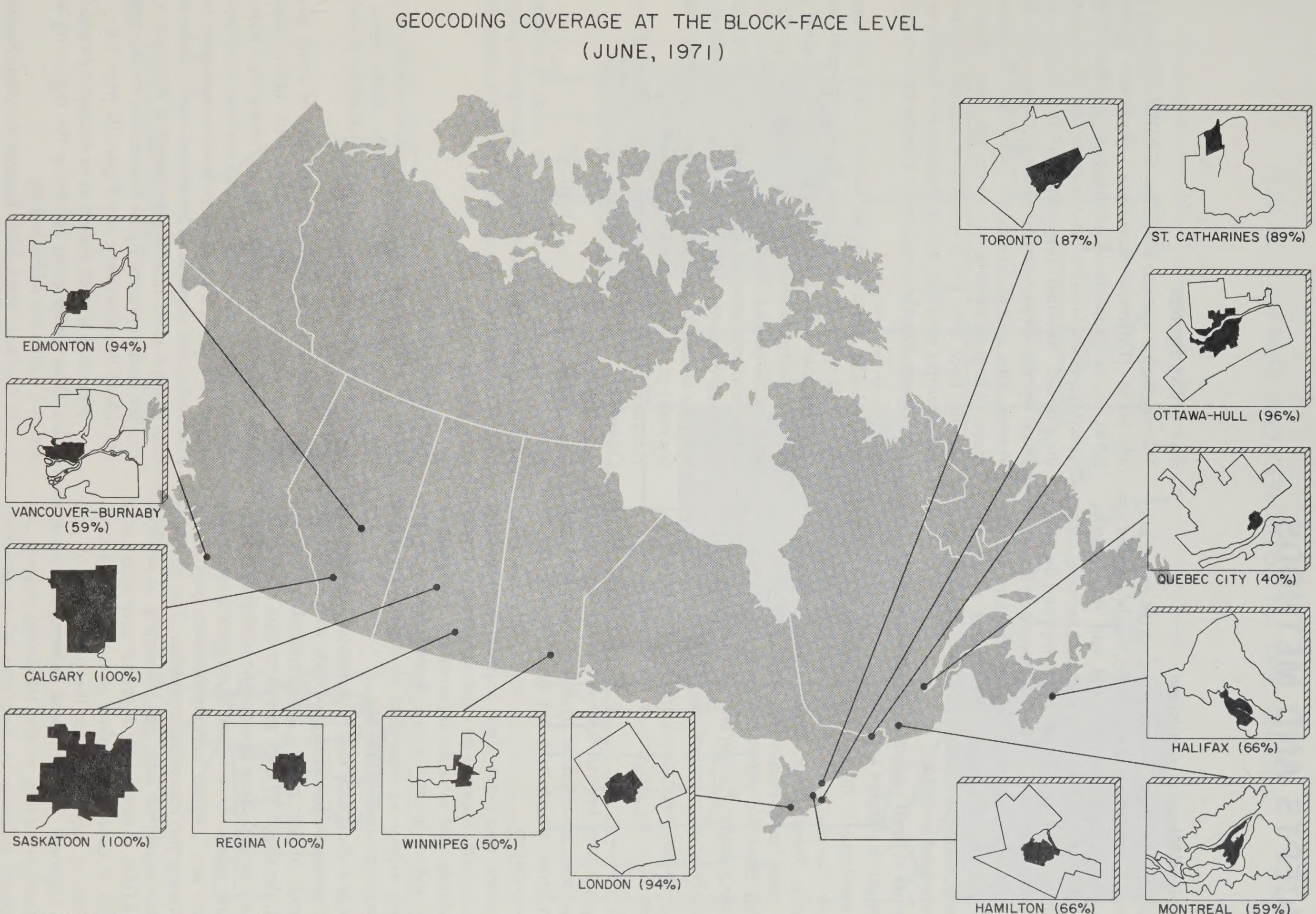
Health Services Planning: A Potential GRDSR Application

A number of factors influence the choice of location for a new hospital or health services clinic in a major city, such as accessibility through major traffic arteries, availability of professional staff, areas most in need of services.

GRDSR can be particularly helpful in deciding which city areas are most in need of proximal medical facilities. One approach is to find out where past patients have lived and what medical services and equipment they required, using city hospital records.

Hospital visitation records bear, in addition to medical content, an address identifier for every patient. Therefore geocoding operations can, in most cases, be carried out on the visitation records. Through GRDSR, considerable statistical information can then be generated (for instance, the incidence of hospital visits originating from each and every portion of the city). The retrieved information can be cross-tabulated by the type of medical services required or by any other item of information contained in the original records. For example, the incidence of various diseases, illnesses, or special health problems in certain city areas can be ascertained. Such statistics can prove to be an invaluable aid in determining which city areas would best be served by neighbourhood medical services or a new hospital.

FIGURE - II



IN EACH DIAGRAM, THE OUTLINE DEPICTS THE BOUNDARY OF THE CENSUS METROPOLITAN AREA WHILE THE SHADED PORTION REPRESENTS THE AREA OF BLOCK-FACE COVERAGE. THE PERCENTAGE OF POPULATION RESIDING IN THE SHADED AREA IS ALSO SHOWN. DIAGRAMS ARE ALIGNED IN THE NORTH-SOUTH DIRECTION.
APPROXIMATELY 7 MILLION, OR 34% OF THE POPULATION OF CANADA ARE NOW COVERED AT THE BLOCK-FACE LEVEL.

CONCEPTS AND METHODS

Review of small-area problems

An urban planner, faced with comparing the expropriation costs of several expressway routes, might attempt to use municipal assessment files and find that records were identified by address, city wards or in some other way. To obtain statistics about land values and dwelling types, the file must be inspected one record at a time to determine which data to include in estimates for the proposed expropriation area. The expense of this approach has been prohibitive but, until recently, few alternatives were available.

Another type of requirement, now directed to the census, might be phrased as follows:

> *"A tabulation of the number of people resident in the Toronto area bordered by Summerhill Avenue, Yonge Street, Mount Pleasant Cemetery, and the boundary for East York is required. Break this tabulation down by age, sex, income, country of origin and occupation."*

Alternatively, another request might read:

> *"Provide the same statistics for the area named Ward Five, as outlined on the attached city map."*

Such requests have been difficult to service, since census data have been summarized by census tracts and enumeration areas, which may not coincide with the required boundaries.

To solve small-area data requests economically, Statistics Canada required an efficient system to repeatedly assemble and tabulate information according to arbitrary special-interest areas. Before describing the conceptual aspects in detail, let us expand upon the operational steps in GRDSR.

Before a data base can be geocoded, each record must be assigned some reference code which identifies the record to its proper geographical source. In GRDSR, the source of each record or data observation is precisely located using a comprehensive geographical coordinate system. Each record is assigned a coordinate value, or "key", which actually becomes part of the record during the geocoding operation. The geocoded file is then stored for later use. Ultimately, at retrieval time, GRDSR automatically identifies each query area with a list of coordinate values and, using the coordinate values as keys, retrieves the precise set of data records required. The retrieved information is then summarized according to the tabulation request; the user receives statistics in the form of a convenient report.

The UTM System

The coordinate system chosen for GRDSR is known as the UTM (Universal Transverse Mercator) System. UTM is an established international convention for specifying point-locations on the globe, and is shown on the popular National Topographical Map Series produced by the Department of Energy, Mines and Resources. This system divides the globe into 60 vertical zones. Altogether, 16 zones cover the land mass of Canada. Each zone has a width of 6 degrees longitude and a central meridian which becomes the vertical axis for the zone. The horizontal axis is formed by the earth's equator.

In UTM, point-locations within a zone are based on two distances in metres (one easting, one northing) from the zone axes. The central meridian is assigned an artificial value of 500,000 metres easting; the equator is assigned the value 0. Distances are measured on a plane rectangular grid onto which the zone's surface features have been projected. The two values are combined with a zone number to arrive at a unique coördinate value for every point on the land mass of Canada.

For example, the UTM coordinates of the Peace Tower, Ottawa, are:

Zone	X	Y
18	445177	5030250

In this way, the UTM coordinates seem to define a point-location to the nearest metre, although the projection of the earth's surface onto a plane grid introduces minor distortions.

Basic definitions

Points at which streets intersect or curve sharply in the city pattern are referred to as *nodes*. Every street is represented by a series of nodes connected by straight-line *segments*.

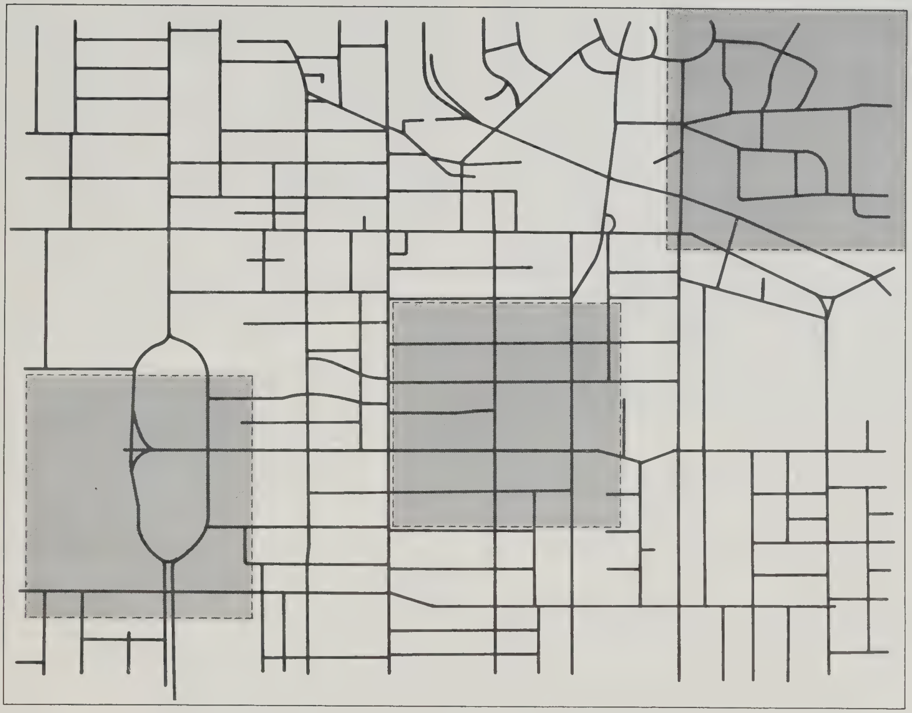
A *block-face* is defined as one side of a city street between consecutive intersections with other streets. Thus, up to two block-faces can be formed by a pair of adjacent nodes, each located at a four-way street intersection. However, a block-face can also encompass several nodes. For example, a block-face may contain one intermediate node marking a change in direction and another node representing an intersection on the opposite side of the street only (see Figure III, page 7).

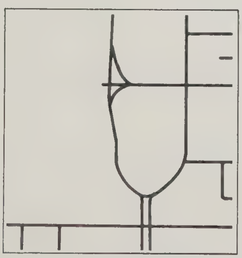
Whenever a block-face is to be formed by a pair of nodes, these nodes must constitute the beginning and end of a valid civic address range.

In this way, block-faces become the basic building blocks used in the GRDSR System.

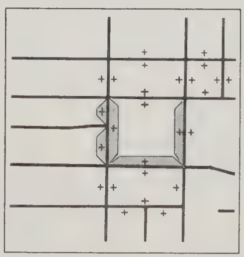
FIGURE III

HOW BLOCK-FACES AND CENTROIDS (+) ARE CHOSEN

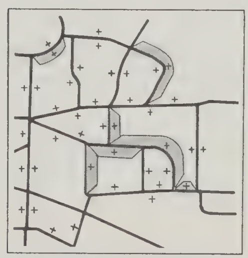




(1) STREETS THROUGH RECREATIONAL OR PARKLAND--NO CENTROIDS



(2) STREETS IN REGULAR (GRID) PATTERN



(3) STREETS IN IRREGULAR PATTERN

THE SHAPES OF SEVERAL BLOCK-FACES ARE SHOWN BY SHADED AREAS

Why addresses are necessary

The GRDSR System is partly based on the premise that most agency records and survey responses are identified, geographically, by the addresses of respondents. An address is the starting point in coordinate assignment, because every street address in an urban centre can be identified as belonging to some block-face.

How addresses are converted into coordinates

In GRDSR, all street addresses along a block-face are assigned, and share, the coordinates of the block-face centroid, which is simply a reference point offset from the street midway between the two nodes forming the block-face. During the conversion operation, the address of each record or data observation is matched to a block-face (using a list of valid street names and address ranges). From there, the correct centroid is known and its coordinates can be added to the record.

The Area Master File

The actual geocoding operation (or assignment of coordinates to data) is carried out using GRDSR components known as Area Master Files (AMF), which will be described in detail in Features and Components, page 10.

Area Master Files contain a logical representation of all city streets, plus some other features, in computer-readable form. An AMF references every street, address range, block-face and centroid coordinate in the covered area. Also itemized are other features (such as railroad tracks, rivers, and municipal boundaries), which help users to choose query areas. During the geocoding operation, centroids are obtained by matching addresses against street names and address ranges within the Area Master File. (In this way, address ranges can be thought of as representing the actual building blocks, rather than block-faces.)

Area Master Files have been created for major portions of 14 Canadian urban centres, which include a total of 16 cities (see Figure II, page 5). These files reference more than 225,000 block-faces, corresponding to a population figure of approximately seven million.

Rural Geocoding Coverage (1971 Census)

The 14 Area Master Files have already served to geocode certain urban portions in the 1971 Census. For the remainder of Canada not covered at the block-face level (urban and rural), census geocoding, as already noted, has been carried out using standard enumeration areas, with one centroid assigned to the approximate population centre of each. Enumeration Areas outside Area Master File coverage number more than 27,000. Tabulation requests for query zones in rural areas or in the urban shadow of developing urban areas are easily (and automatically) handled using centroids at the EA level, the block-face level, or both. During retrieval, inaccuracy in data selection at the EA level is minimized by the choice of centroids near the population centre and by a process of compensation, whereby errors from including or missing centroids in a query area are self-cancelling.

ADVANTAGES, LIMITATIONS OF CONCEPTS

Choice of block-faces

Block-faces become the finest level of resolution possible when each cluster of data observations or survey records belonging to one block-face share the same centroid coordinate. This is a logical outcome of the building block principle adopted by GRDSR.

As one alternative, geocoding to the land parcel or household level achieves higher resolution which may be desirable for some purposes. This approach requires extensive local research. Since block-face resolution is expected to satisfy the vast majority of geocoding requests, land-parcel geocoding could not be justified for a Canada-wide system such as GRDSR.

A second alternative was to identify data by city block, a poorer resolution. However, this approach would not have allowed users enough flexibility, since the integrity of city blocks would have to be respected in specifying query areas. For instance, it would not be possible to obtain tabulations for one side (or both sides) of a city street.

The choice of block-faces as basic geocoding building blocks has several implications. All observations originating from one block-face bear the coordinate of its declared centroid. As a result, the integrity of block-faces should preferably be respected in specifying query areas for retrieval. They should not be split: observations referenced to a split block-face will appear in the results only if the query area includes the block-face centroid. If not, the observations are missed entirely. Another implication is that geocoding to the household or land parcel level (each individual property bears a centroid) is not possible using this system. This may pose definite restrictions on municipal services, engineering and land-banking applications where higher resolution is required.

Identification by street address

Statistics Canada recognizes that a majority of statistical surveys and agency records are address-identified and provides for this with a System component known as the Postal Address Analysis System (PAAS).

In geocoding a file, addresses are analyzed and converted to centroid co-ordinates. Because the conversion is done by computer, complete addresses must be decomposed into separate, clearly identified components (such as street name, type, house number and municipality name). Because PAAS achieves a high efficiency and success rate, address specifications of relatively poor quality can still be geocoded. This feature clearly extends the scope of GRDSR applications. More information about PAAS is provided in Features and Components, page 10.

However, GRDSR cannot perform the geocoding operation on records which, by their nature, are not identified to street addresses. Certain city facilities, such as sewers, gas and hydro lines, traffic signals and overhead structures may be of interest from a geocoding standpoint. In this case, the user must geocode the file before submitting it to GRDSR.

Choice of coordinate system

While UTM is ideally suited to geocoding at the block-face level it has some limitations in land survey and civil engineering operations where the 3 Transverse Mercator System is more accurate, and thus a frequent choice. However, programs are available to convert files geocoded with the UTM system to 3 TM and vice-versa.

FEATURES AND COMPONENTS

The Area Master File

How an AMF is created

Geocoding starts with an accurate street map. A large-scale, current map showing block-face address ranges is required, together with an up-to-date street index. After the map is divided into sections a node is assigned to each street intersection. Nodes are also assigned to points where streets begin, end or curve sharply. A non-distorting overlay is prepared for each map section and the position of each node is marked on the overlay.

Once serial numbers have been assigned to the nodes, descriptive codes for every street segment are transcribed onto a specially-prepared form. The codes include feature names, types, directions, node numbers, and addresses at the intersections. Then the overlay is placed on a digitizing table. The digitizing equipment measures node positions relative to control points on the overlay, and generates one horizontal and vertical "table" coordinate for each node. Since the UTM coordinates of the control points are known in advance, the UTM coordinates for the nodes can then be calculated from the table coordinates. During subsequent computer processing, centroid coordinates are calculated for each block-face using the coordinates of the two nodes bordering the block-face. Finally all items are merged to create an Area Master File for the city (see Figure IVa, page 11).

How the AMF is used

Three operations, each related to address conversion, require files of information contained in the Area Master File. To eliminate the maintenance and updating of three separate files, each is derived from a clean, up-to-date AMF as required.

(i) Street name lists are used by PAAS to verify input addresses prior to the assignment stage.

(ii) The Address Conversion File (ACF) is used to obtain centroid coordinates for input addresses once the PAAS stage is complete. Addresses are matched against block-face address ranges and the corresponding centroid coordinates are selected from the ACF. Geocoding is complete once centroid coordinates replace addresses in the original file.

(iii) The Block-Face File was created specifically to geocode the 1971 Census. This file makes it possible to link parcels of census data, which are not otherwise address-identified, to block-faces and centroids.

As a geographic base file, the Area Master File design is unique. The central concept is to provide a geographical framework that is as practical as possible for a variety of potential users, but efficient from a file creation/update standpoint.

A series of error-handling and correction procedures comes into play whenever Area Master Files are being built or updated. Extensive computer checking is done to ensure that each node is linked to the correct street segments, and vice-versa. This process locates the majority of clerical errors. When each section file is complete, it is plotted at the same scale as the original map. The two maps are then compared to verify node locations. Usually, further plotting followed by two to three update cycles, will produce a clean Area Master File.

Local area breakdowns, such as census tracts, electoral wards, city wards, and other extra codes were purposely excluded from the AMF. Its design is such that these jurisdictions are easily constructed independently of the AMF, but using the identical building-block technique. Because areal boundaries are constantly changing, their inclusion would have seriously prolonged the operation needed to build and maintain an accurate, up-to-date base file.

Urban Street Maps

Computer-plotted street maps are an important by-product of building an Area Master File. Because the AMF is a logical representation of city features, its contents can be used, in reverse, to create facsimile maps at any scale. Plotting is accomplished using the GRDSR component, MAPMAKR (see Figure V, page 15).

These maps have several purposes:

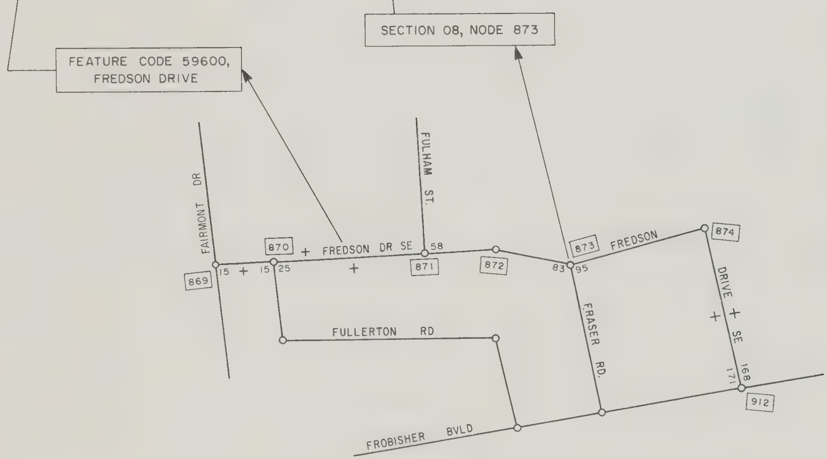
- The best way to edit or validate an Area Master File is to recreate the original map, using the plotter. Errors and inconsistencies are clearly highlighted.
- They provide a return service to municipalities who in turn are aware of what updates are required.
- The maps are supplied to users for outlining query areas and depict city features as seen by the AMF.

Using MAPMAKR, maps can be produced to suit a variety of purposes. It is possible to pre-specify the area to be plotted and the scale required. Parameters are used to determine whether various options, such as nodes, feature names, centroids, address ranges and control points, will appear on the final plot.

FIGURE - IV a

THE AREA MASTER FILE (SPECIAL FORMAT)

MUNIC CODE	FEATURE CODE	SEQ. No.	STREET NAME	TYPE	DIR.	NODE NUMBER	NODE COORD. X	NODE COORD. Y	ADDRESS BEFORE L	ADDRESS BEFORE R	ADDRESS AFTER L	ADDRESS AFTER R	CENTROID LEFT X	CENTROID LEFT Y	CENTROID RIGHT X	CENTROID RIGHT Y	INTERSECTING FEATURES
4835	59600		FREDSON	DR	SE												
		040E				08912	706954	5651540	168	171	—	—	706868	5651755	706861	5651673	FROBISHER BV
		035				08874	706844	5651755	—	—	—	—	—	—	—	—	
		030				08873	706771	5651722	—	83	—	95	—	—	706635	5651700	FRASER RD
		025				08872	706733	5651729	—	—	—	—	—	—	—	—	
		020				08871	706631	5651722	—	—	58	—	706516	5651734	—	—	FULHAM ST
		015				08870	706497	5651711	—	15	—	25	—	—	706452	5651687	FULLERTON RD
		010B				08869	706406	5651707	—	—	—	15	—	—	—	—	FAIRMONT DR



(+) CENTROIDS FOR FREDSON DR SE

Postal Address Analysis System

Addressing conventions vary according to locality, language and post office regulations, but few comprehensive systems are available to digest and organize a file of street addresses. The PAAS system is a flexible and inexpensive device for accomplishing this job. For geocoding applications, addresses can originate from any city having an Area Master File at Statistics Canada (see Figure II, page 5). Otherwise, PAAS can re-structure and organize virtually any address file in use today.

While the number of addressing conventions across Canada is considerable and many conventions often appear in one file, PAAS consistently demonstrates a high success rate at exceptionally low cost. In its current version, it accepts street addresses (including municipality names) in completely free format and decomposes each address into several elements such as street name, street type and direction. The addresses are then matched against a subset of the Area Master File (the Address Conversion File) and, if the match is successful, a centroid coordinate is assigned to each record in the original file (see Figure IVb, page 13).

The flexibility of PAAS is enhanced through parameters which are passed to the program when geocoding starts. These parameters improve PAAS efficiency by indicating the nature and characteristics of the incoming addresses.

Significantly, the entire conversion process is accomplished at an average cost of less than one cent per address.

The Query Area Library

Many users are expected to submit special-purpose areas for data retrieval and refer to them repeatedly in making requests. Statistics Canada also expects continuing requests for census statistics arranged by the traditional standard areas — provinces, counties, census tracts and enumeration areas. (Altogether, there are 13 distinct sets of standard census areas, each set covering most of the settled area of Canada. The 13 sets comprise more than 53,000 separate areal units.)

Before information about any query area can be retrieved, GRDSR must define the query area in terms of the geocoded data base. Definition is accomplished by associating the area name with "pointers", which indicate precisely where the desired elements can be found. Pointer sets for each standard census area and for special-purpose areas are kept in a system component called the Query Area Library. A QAL is opened specifically for each new data file stored in GRDSR.

In normal practice, users outline the boundary of a special query area on a map. Vertices along the boundary are located using a digitizer so that their positions can be converted to UTM coordinates. A computer-programmed algorithm is used to test whether each successive centroid coordinate in the data base belongs in the query area. Finally, the coordinates selected are converted to pointers, which serve to locate the corresponding data elements required.

To avoid repeating this process, frequently-used area names are stored in the Query Area Library. Each area name is associated with a set of pointers. Areas that will be requested often and by different users are stored in a portion of the QAL reserved for permanent areas. Other area definitions will be stored for a limited time in the temporary QAL. Several other methods for designating query areas are described in Operations, page 18.

STATPAK

STATPAK was developed for GRDSR as a generalized program to retrieve statistics efficiently by arbitrary areas. Users communicate with STATPAK through the problem-oriented language TARELA and receive statistics in the form of convenient, easy-to-read tables.

How a file is stored

STATPAK's efficiency is made possible by changing the structure of an incoming file after it has been geocoded. Instead of keeping all data characteristics for a respondent together in one record, each data characteristic is handled separately. The entire set of responses for one data characteristic are assembled and stored as a continuous string. Because more than one record is usually attached to each centroid coordinate, an index is built to locate precisely where the responses for each distinct coordinate value are found. The index is then used to provide pointers for new query areas before they enter the QAL.

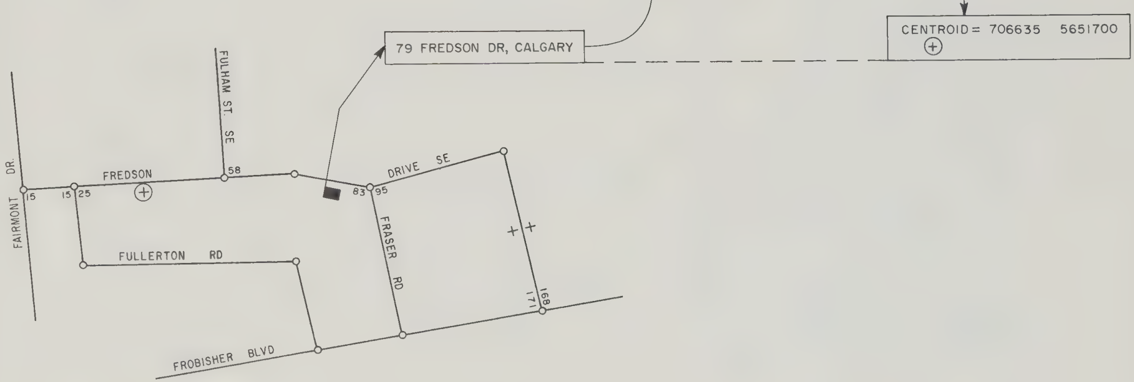
Complete strings of data characteristics are finally stored on direct-access devices and a Query Area Library is established for the file.

To visualize the final geocoded data base, imagine a huge matrix. Using the 1971 Census as an example, 21.6 million people counted in the census are arranged vertically in order of their centroid coordinates along the left side. Approximately 120 data characteristics form vertical parallel strings suspended from the top of the matrix. Instead of storing all characteristics for a person in a self-contained record, one string is created for each characteristic (such as age, sex, marital status, income or occupation). This method allows each string to be compressed to occupy the least possible space for

FIGURE- IV b

THE ADDRESS CONVERSION FILE (SPECIAL FORMAT)

STREET NAME	TYPE	DIR	FEATURE No.	ODD/ EVEN	ADDRESS RANGE LOW	ADDRESS RANGE HIGH	CENTROID ZONE	CENTROID X	CENTROID Y
FRASER									
FREDSON	DR	SE	59600	ODD	15	15	18	706452	5651687
"	"	"	"	ODD	25	83	18	706635	5651700
"	"	"	"	ODD	95	171	18	706861	5651673
FREDSON	DR	SE	59600	EVEN	58	168	18	706868	5651755
FROBISHER									



CENTROIDS (+)

the information contained. As a result, the use of costly direct-access storage space is minimized.

During the storage operation a name for each data characteristic is retained along with code names for the values the characteristic can assume. The names appear in a document called the Data Dictionary which is used, in turn, to code TARELA requests. The problem-oriented nature of TARELA rests on these names, because they are chosen by subject-matter specialists when files are submitted to GRDSR for geocoding.

How information is retrieved

After STATPAK accepts and analyzes a TARELA request, it generates a tailor-made program to retrieve the data. The program is then executed.

The operating advantage rests on large files where only a small portion is accessed at one time, that is, whenever tabulations are requested for small areas or relatively few data characteristics. Because data are retrieved in direct-access mode, the actual execution cost is strictly dependent on the extent of the query area and on the nature of the tabulations required, not on the size of the whole file. In a file of 1.5 million records, the costs of a tabulation vary from $30 to over $100, depending on the number of records to be retrieved.

Any file which has fixed or variable-length records with geographic identification (ideally UTM coordinates) may be reorganized into a form acceptable to STATPAK. Written in PL/1, STATPAK is a set of modules assembled into a tailor-made source program for each new TARELA request. The tailor-made program is exceptionally efficient because it immediately locates the required data string and accesses only those portions belonging to the user-defined query area. It is erased when the final tabulation is complete.

STATPAK is implemented at Statistics Canada on the System/360-65, under OS/MVT and HASP, and occupies roughly 150 K bytes of core storage.

TARELA

Tabulation requests are coded in a highly user-oriented language called TARELA, requests in which are submitted directly to the system and will normally be returned within one or two days, depending on the computer workload and the size of tabulations requested.

As a data retrieval language TARELA offers significant advantages. It spares non-programming users the trouble of writing retrieval requests for subsequent analysis and programming, and it frees programmers and analysts for more complex work, such as refining the GRDSR System. Programming, debugging and testing delays are bypassed. Finally, potential communications problems inherent in dealing with different professional groups are avoided because the ultimate user can himself communicate directly with the data base.

To write a TARELA request, users must have access to the appropriate Data Dictionary created when their data base was geocoded. A standard data dictionary for the census files will be available to interested users. Using the dictionary, each response characteristic (i.e. age, sex, occupation) is selected by name and code words representing numerical values appearing in the data base. The user can also specify appropriate functions (such as a COUNT of persons satisfying some criteria, or SUM and AVERAGE of a set of retrieved data values). As the request is formulated, coded information is simply written after each TARELA keyword as shown in Figure VI, page 17.

Data Mapping by Computer

MAPPAK is a facility to display spatial distributions of a statistic in the form of a map. MAPPAK operates as an interface between STATPAK and SYMAP, a mapping program developed at the Laboratory for Computer Graphics, Harvard University.

Reading statistical data from a map often has compelling advantages over having the same information tabulated in report form. Inspection of the map can instantly show where extreme values of some function occur. A map can highlight problem areas at a glance.

MAPPAK can be used to stratify data values into several classes or to filter a data characteristic. The results are shown as numbers or as shaded areas on the paper surface.

For instance, if a MAPPAK user is interested in census data, the distribution of average income can be depicted in many levels of shading over a city area. Or, a user can specify that an area be subdivided into 400 by 800 foot rectangular cells, with the average number of children per family shown as a number within each cell. To illustrate the filtering characteristic, MAPPAK can be requested to shade city areas where half the population is of foreign origin, or where a majority of families rent rather than own homes.

The uniform data areas generated by MAPPAK can take many forms. Users can request data relative to any grid cell pattern, by rectangles of any size, or in terms of concentric circles. At one extreme, a data value can be mapped for every centroid point in the city area (subject to confidentiality constraints). At the other extreme, a single data value for some arbitrary area sketched on a street map can be obtained.

FIGURE-V

COMPUTER-PLOTTED STREET MAP OF SASKATOON (PORTION)

MAPPAK incorporates all SYMAP facilities including contour mapping of surface data, classification of data values within arbitrary, pre-defined areas and summing the distribution of a set of data values. It has the flexibility to display detail down to the finest level on the data base (the block-face) and can generate maps to any desired scale (see Figure VII, page 19).

Again, it must be pointed out that the routines for confidentiality checking will be applied when MAPPAK is used to retrieve data from the geocoded census files. The routines will operate in the same manner as for regular statistical tabulations.

FIGURE-VI

PREPARING A TARELA REQUEST

```
RESEARCH PROJECT PLANNING ZONES

AREA SET ID--
TS01674000

AREA ID---     DESCRIPTION ---

TA01674001     PLANNING ZONE 1
TA01674002     PLANNING ZONE 2
TA01674003     PLANNING ZONE 3
TA01674004     PLANNING ZONE 4
TA01674005     PLANNING ZONE 5
```

```
TARELA DATA DICTIONARY FOR FILE
****   GEOCODED 71 CENSUS FILE SHORT FORM
FILENAME:GE71CFSF;

PART3 : VARIABLE AND CODE DESCRIPTION
VARIABLES ON LEVEL POPULATION
```

VARIABLE INTERPRETATION	VARIABLE NAME	CODE OR CODE RANGE	CODE NAME	CODE INTERPRETATION	STUB TEXT
RELATIONSHIP TO HEAD OF HOUSEHOLD	RELTOHD	1	HEAD_OF_HOUSEHLD	HEAD OF HOUSE HOLD	HEAD OF H(
		2	WIFE	WIFE OF HEAD OF HOUSEHOLD	WIFE OF HE
		3	SON	SON OF HEAD OF HOUSEHOLD	SON OF HE/
		4	DAUGHT_IN_LAW	DAUGHTER-IN-LAW OF HEAD OF HOUSEHOLD	DAUGH,TER-
		5	SON_IN_LAW	SON-IN-LAW OF HEAD OF HOUSEHOLD	SON-,IN-,l
		6	DAUGHTER	DAUGHTER OF HEAD OF HOUSEHOLD	DAUGH,TER
		7	GRANDSON	GRANDSON OF HEAD OF HOUSEHOLD	GRAND,SON
		8	GRANDDAUGHTER	GRANDDAUGHTER OF HEAD OF HOUSEHOLD	GRAND,DAU(
TOTAL INCOME OF PERSON	TOTINCOM	0-99999		TOTAL INCOME IN $	

```
FILE: GE71CFSF;
AREA: TA01674001;
DEFINE: HEAD_INC FØR HØUSEHØLD AS TØTINCØM IF RELTØHD=1;
HEADING: NUMBER ØF HØUSEHØLDS
            AVERAGE INCØME ØF HEAD ØF HØUSEHØLD
            EACH BY NUMBER ØF RØØMS IN DWELLING AND NUMBER ØF PERSØNS
            IN HØUSEHØLD;
CHARACTERISTICS:
     PERSINHH 1,2,3,4,5 ØR MØRE;
     RØØMS 1, 2,3,4-5, 6 ØR MØRE;
TABULATE:
     CØUNT,
     AVERAGE(HEAD_INC)  /*AVERAGE HEAD INCØME*/;
```

NUMBER OF HOUSEHOLDS
AVERAGE INCOME OF HEAD OF HOUSEHOLD
EACH BY NUMBER OF ROOMS IN DWELLING AND NUMBER OF PERSONS
IN HOUSEHOLD
2 FEBRUARY 1972 PAGE 1

AREA: PLANNING ZONE 1

	ROOMS 1	ROOMS 2	ROOMS 3	ROOMS 4-5	ROOMS 6 OR MORE
PERSINHH 1					
COUNT...............	15	15	40	40	30
AVERAGE HEAD INCOME.	3,555	3,035	1,810	1,515	850
PERSINHH 2					
COUNT...............	5	15	40	180	115
AVERAGE HEAD INCOME.	3,010	3,020	2,725	3,250	3,500
PERSINHH 3					
COUNT...............	-	5	15	140	75
AVERAGE HEAD INCOME.	-	3,050	3,615	4,380	3,990
PERSINHH 4					
COUNT...............	5	-	5	135	90
AVERAGE HEAD INCOME.	2,550	-	3,305	4,835	4,910
PERSINHH 5 OR MORE					
COUNT...............	-	-	10	200	285
AVERAGE HEAD INCOME.	-	-	3,190	5,350	5,150

IN THE REQUEST, THE AREA ID IS PROVIDED BY THE AREA LIST (TOP LEFT). THE FILE ID, VARIABLE NAMES AND CODE NAMES ARE PROVIDED BY THE DATA DICTIONARY SHOWN (IN PART) AT THE TOP RIGHT. IN THE OUTPUT TABULATION (ABOVE) THE AREA DESCRIPTION, "PLANNING ZONE 1" HAS BEEN RETRIEVED FROM THE QUERY AREA LIBRARY.

OPERATIONS

Handling User Surveys

Many geocoding applications are of interest to municipal administrations. GRDSR can be used to access information of significant importance to urban planning and administrative processes. Several possible applications were described on page 4.

Geocoding and Data Storage

The geocoding operation can now be carried out in 14 larger Canadian urban centres having Area Master Files at Statistics Canada (see Figure II, page 5). Since GRDSR is fully computerized, input files can be in machine-readable form (such as punched cards or magnetic tape). Any unusual address structures in the input file may require definition prior to submission. Once the user has provided a description of the file (including record length, variable names and values, address location, etc.) geocoding can begin. This can happen in one of two ways. Users in the public sector (municipal, provincial, and federal governments) who have suitable computing facilities may obtain the GRDSR System for their own use. In other cases this operation as well as subsequent data retrievals may be carried out by Statistics Canada under contract.

In either case the actual processing phases are as follows:

- To geocode the input file, address identifiers are removed, analyzed and used to assign a centroid coordinate to each record.
- Data characteristics are gathered together and arranged in strings. The re-organized file, together with control information describing the strings, is stored on direct-access devices, ready for data retrieval.
- A Query Area Library is opened for the file.
- Finally, a Data Dictionary for preparing TARELA requests is created.

Data Retrieval

Any number of retrievals can be carried out once the storage operation is complete. Definition of query areas for the retrieval phase can also be done in several ways. Initially, it will be necessary to outline the desired query areas on a city map, name them clearly, and submit these specifications along with the tabulation request. For instance, query areas for a municipal retrieval might be defined as "Planning Zones I, II, III, and IV". Mapped query area boundaries are then digitized and converted to UTM. At this point, the area definitions are stored in the Query Area Library opened specifically for this file. Subsequent tabulation requests for these areas can then be referred to the QAL for definition rather than repeating the UTM conversion operation.

From this point onwards, it is possible to obtain data tabulations through the GRDSR System. TARELA is used as a vehicle for the tabulation request, which is coded using the Data Dictionary. For instance, if the input data base was an assessment file, a tabulation of assessed value for various dwelling types could be specified using convenient characteristic names (such as "VALUE", "DWELLTYPE"), area names (such as "PLANNINGZONE4"), file names (such as "ASSMFIL6") and parameters indicating the format of tabulations desired.

Requests for further tabulations can be handled in a similar manner. Tabulation requests can be processed with exceptionally quick turnaround once files have been geocoded.

How users can specify areas

An important feature of GRDSR is that it accepts area descriptions in several convenient ways.

Outlines on maps

Users will probably find it most convenient to outline query areas on a map. In most cases any convenient map can be chosen. Statistics Canada is producing copies of computer-plotted city maps, which are particularly appropriate for graphically displaying block-faces in each city area. Outside the urban areas of block-face geocoding coverage, users will be advised to use the National Topographic Series (NTS) maps produced by the Department of Energy, Mines and Resources. The important thing is that users choose an appropriate map scale, then mark out and name query areas as clearly and as accurately as the problem demands.

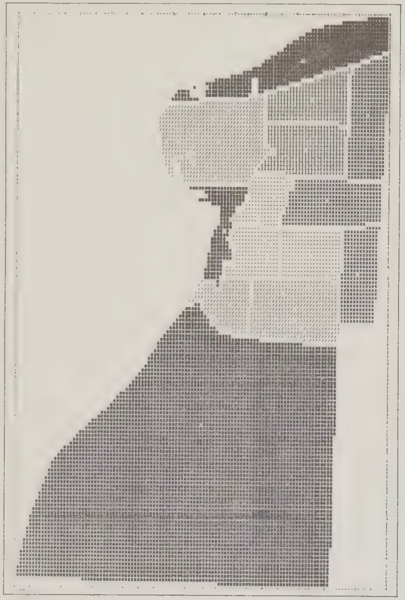
Defined by features

It will be possible to specify a query area in terms of known features (rivers, streets, railroad tracks). For instance, an Ottawa user could describe in writing that Research District No. 5 consists of an area bounded on the north by St. Patrick Street, on the east by Chapel Street, on the south by Templeton Street and on the west by Nelson Street. It is possible to request data for street-oriented query areas in the same manner. For example, a user could request statistics for one side of Rideau Street, in the municipality of Ottawa; from Sussex Drive to King Edward Avenue (odd-numbered side, even-numbered side, or both).

Rather than submit a list of feature names, users can define areas by a set of node numbers copied from the computer-plotted city maps. Nodes are chosen at

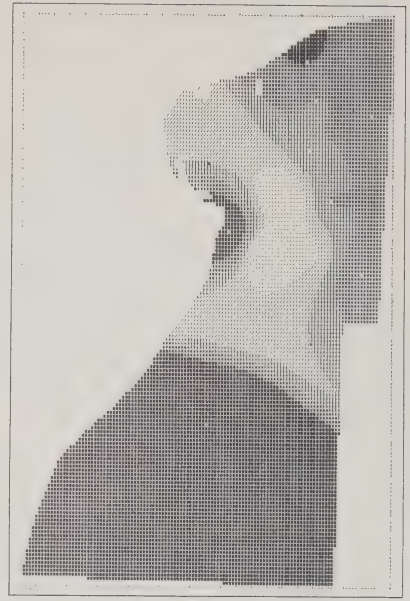
FIGURE VII

MALE/FEMALE RATIO FOR THE CITY OF SARNIA



(1) CONFORMAL

(2) PROXIMAL



(3) CONTOUR

MAPS ARE ALIGNED IN THE NORTH-SOUTH DIRECTION. BOUNDARIES FOR THE CONFORMAL MAP ARE FORMED BY CENSUS TRACTS. FOR THE OTHER TWO MAPS, BOUNDARIES ARE DETERMINED FROM THE DATA. RATIO VALUES ARE DIVIDED INTO FIVE CLASS INTERVALS BETWEEN EXTREME VALUES OF 0.89 AND 1.18. EACH HIGHER CLASS INTERVAL IS REPRESENTED BY A PROGRESSIVELY DARKER SHADE.

points where the boundary features intersect. Thus, the area perimeter is defined by the nodes, which are matched to the Area Master File before storing the area in the QAL.

Using grid coordinates

UTM coordinates can be used to specify query areas in two ways. Data can be retrieved according to a list of individual centroids chosen from the Area Master File. Or, a set of coordinates along a boundary can be used by the system to calculate an enclosed area.

Using area names

Once an area has been submitted using one of the above methods, its name and description are entered and stored, temporarily, in the Query Area Library. For subsequent references the QAL description will be referenced directly by area name, bypassing the map conversion operation.

Of course, all requests for census data by traditional standard areas will also be serviced through the Query Area Library. The QAL contains a pointer set for each province, county, municipality, census tract, enumeration area, and all other standard geostatistical areas used in the 1971 census.

Using other areas

The system permits addition and subtraction of query areas to form a new query area. For instance, a user can outline and request statistics for six areas on a map, naming these areas Area 1, Area 2, ... Area 6. He can then request further data for a new zone, defined as follows:

QZONE1 = Area 1 + Area 2 + Area 3

If Area 5 is contained within Area 6, the following specification would result in a doughnut-shaped query zone:

QZONE2 = Area 6 - Area 5

FURTHER INFORMATION

Users who are primarily interested in census statistics using GRDSR may obtain further information by contacting:

User Inquiry Service

Census Division

Statistics Canada

Ottawa, K1A 0T6

Statistics Canada is prepared to provide assistance and further information to users who wish to geocode their own data files. Detailed system documentation will be available in response to technical requests. This information will be provided by a manual entitled *A Technical Description of the GRDSR System*, followed by User Manuals for certain components. For further information of a specialized or technical nature, please contact:

General Survey Systems
Methodology and Systems Branch
Statistics Canada
Ottawa, K1A 0T6

966-6586